Data Analysis and Visualization in R (IN2339)

Exercise Session 4 - Low dimensional visualization

Daniela Klaproth-Andrade, Jun Cheng, Daniel Bader, Julien Gagneur

Section 00 - Getting ready

1. Make sure you have already installed and loaded the following libraries:

```r
library(ggplot2)
library(data.table)
library(magrittr)   # Needed for %>% operator
library(tidyr)

library(MAS6005)  # Install with devtools::install_github("OakleyJ/MAS6005")
library(ggrepel)
```

Section 01 - Plot types

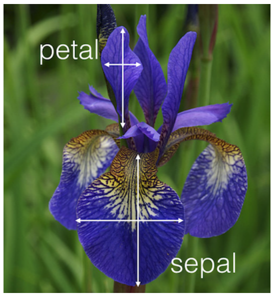
Match each chart type with the relationship it shows best.

1. shows distribution and quantiles, especially useful when comparing distributions.
2. highlights individual values, supports comparison and can show rankings or deviations categories and totals
3. shows overall changes and patterns, usually over intervals of time
4. shows relationship between two continues variables.

Options: bar chart, line chart, scatterplot, boxplot

Section 02 - Visualizing distributions

`Iris` is a classical dataset in machine learning literature. It was first introduced by R.A. Fisher in his 1936 paper.



1. Load the *iris* data and transform it to a `data.table`. Have a look at its first and last rows.

2. How are the lengths and widths of sepals and petals distributed? Make one plot of the distributions with multiple facets. *Hint:* You will need to reshape your data so that the different measurements (petal

length, sepal length, etc.) are in one column and the values in another. Remember which is the best plot for visualizing distributions.

3. Vary the number of bins in the created histogram. Describe what you see.

4. Visualize the lengths and widths of the sepals and petals from the iris data with boxplots.

5. Add individual data points as dots on the boxplots to visualize all points. Discuss: in this case, why is it not good to visualize the data with boxplots? *Hint:* `geom_jitter()` or `geom_dotplot()` .

6. Alternatives to boxplot are violin plots (`geom_violin()`). Try combining a boxplot with a violinplot to show the the lengths and widths of the sepals and petals from the iris data.

7. Which pattern shows up when moving from boxplot to violin/bean plot? Investigate the dataset to explain this kind of pattern, provide with visualization.

Section 03 - Visualizing relationships

1. Are there any relationships/correlations between petal length and width? How would you show it?

2. [OPTIONAL] Change your plot title and axis labels in the previous plot. For instance, the new title can be "Relationship between petal length and width", and the axis labels "Petal Length" and "Petal Width", respectively.

3. Do petal lengths and widths correlate in every species? Show this with a plot.

Section 04 - The importance of data visualization

Anscombe's quartet was constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it, and the effect of outliers on statistical properties. `anscombe` is directly built in R. You don't need to load it.

1. We reshaped the original `anscombe` data to `anscombe_reshaped`. Which one is tidier?

```
anscombe_reshaped <- anscombe %>%
  as.data.table %>%
  .[, ID := seq(nrow(.))] %>%
  melt(id.var=c("ID")) %>%
  separate(variable, c("xy", "group"), sep=1) %>%
  dcast(... ~ xy) %>%
  .[, group := paste0("dataset_", group)]
```

2. Compute the mean and standard deviation of each variable for each group. What do you see?

3. For each dataset, what is the Pearson correlation between x and y? *Hint:* `cor()` and Wikipedia[1] for Pearson correlation.

4. Only by computing statistics, we could conclude that all 4 datasets have the same data. Now, plot x and y for each dataset and discuss.

5. [OPTIONAL] Consider now the datasets given in the file `boxplots.csv`. Load the data and visualize the different datasets with a boxplot. What do you see? What can you conclude?

6. [OPTIONAL] Exchange the boxplots by violin plots in the previous exercise. Did something change? What do you conclude?

[1] https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Section 05 - Axes scaling and text labeling

1. Consider the `medals` dataset from the `MAS6005` library. Compare total number of medals won against population size in the 2016 Rio olympics with a scatter plot. You can load the dataset with the following code:

```r
library(MAS6005)
attach(medals)
medals_dt <- as.data.table(medals)
```

2. What are the problems with the previous plot? Solve these issues with an adapted version of the plot.

3. Add the country labels to the points in the scatter plot. Compare the differences of using the library `ggplot2` and the library `ggrepel` for this task

Section 06 - Understanding and recreating boxplots

1. [OPTIONAL] Using the `mtcars` dataset, make a boxplot of the miles per gallon (mpg) per cylinder (cyl).

2. [OPTIONAL] Now, recreate the same plot without using `geom_boxplot`. You have to add all the layers manually: IQR box, median line, whiskers and outlier points. *Hint*: Remember how a boxplot is constructed[2]. You may find these functions useful: `IQR`, `geom_crossbar`, `geom_segment`, `geom_point`. Use `data.table` commands.

[2] http://docs.ggplot2.org/current/geom__boxplot.html